Less effective selection leads to larger genomes

Lefébure, Tristan; Morvan, Claire; Malard, Florian; François, Clémentine; Konecny-Dupré, Lara; Guéguen, Laurent; Weiss-Gayet, Michèle; Seguin-Orlando, Andaine; Ermini, Luca; Der Sarkissian, Clio; Charrier, N. Pierre; Eme, David; Mermillod-Blondin, Florian; Duret, Laurent; Vieira, Cristina; Orlando, Ludovic Antoine Alexandre; Douady, Christophe Jean

Published in:
Genome Research

DOI:
10.1101/gr.212589.116

Publication date:
2017

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY-NC

Citation for published version (APA):
Lefébure, T., Morvan, C., Malard, F., François, C., Konecny-Dupré, L., Guéguen, L., ... Douady, C. J. (2017). Less effective selection leads to larger genomes. *Genome Research*, *27*(6), 1016-1028. https://doi.org/10.1101/gr.212589.116

Less effective selection leads to larger genomes

Tristan Lefébure,[1] Claire Morvan,[1] Florian Malard,[1] Clémentine François,[1] Lara Konecny-Dupré,[1] Laurent Guéguen,[2] Michèle Weiss-Gayet,[3] Andaine Seguin-Orlando,[4] Luca Ermini,[4] Clio Der Sarkissian,[4] N. Pierre Charrier,[1] David Eme,[1] Florian Mermillod-Blondin,[1] Laurent Duret,[2] Cristina Vieira,[2,5] Ludovic Orlando,[4,6] and Christophe Jean Douady[1,5]

[1]Université de Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5023, ENTPE, Laboratoire d'Ecologie des Hydrosystèmes Naturels et Anthropisés, F-69622 Villeurbanne, France; [2]Université de Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622 Villeurbanne, France; [3]Université de Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5310, INSERM, Institut NeuroMyoGène, F-69622 Villeurbanne, France; [4]Center for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, 1350K Copenhagen, Denmark; [5]Institut Universitaire de France, F-75005 Paris, France; [6]Université de Toulouse, University Paul Sabatier (UPS), CNRS UMR 5288, Laboratoire AMIS, F-31073 Toulouse, France

The evolutionary origin of the striking genome size variations found in eukaryotes remains enigmatic. The effective size of populations, by controlling selection efficacy, is expected to be a key parameter underlying genome size evolution. However, this hypothesis has proved difficult to investigate using empirical data sets. Here, we tested this hypothesis using 22 de novo transcriptomes and low-coverage genomes of asellid isopods, which represent 11 independent habitat shifts from surface water to resource-poor groundwater. We show that these habitat shifts are associated with higher transcriptome-wide d_N/d_S. After ruling out the role of positive selection and pseudogenization, we show that these transcriptome-wide d_N/d_S increases are the consequence of a reduction in selection efficacy imposed by the smaller effective population size of subterranean species. This reduction is paralleled by an important increase in genome size (25% increase on average), an increase also confirmed in subterranean decapods and mollusks. We also control for an adaptive impact of genome size on life history traits but find no correlation between body size, or growth rate, and genome size. We show instead that the independent increases in genome size measured in subterranean isopods are the direct consequence of increasing invasion rates by repeat elements, which are less efficiently purged out by purifying selection. Contrary to selection efficacy, polymorphism is not correlated to genome size. We propose that recent demographic fluctuations and the difficulty of observing polymorphism variation in polymorphism-poor species can obfuscate the link between effective population size and genome size when polymorphism data are used alone.

[Supplemental material is available for this article.]

Eukaryotic organisms exhibit striking variations in their genome size (GS). Within animals, the range of GS extends from 20 Mb in the roundworm *Pratylenchus coffeae* to 130 Gb in the lungfish *Protopterus aethiopicus* (http://www.genomesize.com/). GS shows no correlation with organism complexity, an observation early on referred to as the C-value paradox (Thomas 1971). Although the contribution of mechanisms such as polyploidization events or transposable element amplification to DNA gain or loss is now better understood (Gregory 2005), the evolutionary origin of GS variation still remains largely unexplained (Petrov 2001).

Large genomes mostly consist of noncoding DNA (Gregory 2005; Lynch 2007). The origins of the large variation in the amount of noncoding DNA found across eukaryotes are currently tentatively explained through two opposite sets of theories. Adaptive theories postulate that variation of the amount of noncoding DNA results in significant phenotypic changes and thus evolves under the control of natural selection. Main examples of phenotypic changes commonly associated to GS variations include nucleus and cellular sizes (Cavalier-Smith 1982), growth rate (Grime and Mowforth 1982), and metabolic rate (Vinogradov 1995) variations. Conversely, nonadaptive theories postulate that GS variations have little phenotypic impact (Doolittle and Sapienza 1980), leaving nonadaptive forces such as mutation and genetic drift as the main evolutionary drivers underlying GS variation (Lynch and Conery 2003). In particular, the mutational-hazard (MH) hypothesis suggests that slightly deleterious mutations, including those that lead to GS variation, can segregate in small populations where the efficacy of purifying selection is impaired by genetic drift (Lynch 2011; Lynch et al. 2011). Under this hypothesis, the evolution of GS would be controlled by the balance between the emergence of large-scale insertions and deletions (indels) and their fixation rate, which ultimately depends on the efficacy of selection and, thus, the effective population size (N_e).

Phylogenetic inertia, varying mutational patterns, and uncertainties in N_e estimates, are but a few difficulties that complicate

Corresponding author: Tristan.Lefebure@univ-lyon1.fr

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.212589.116.

testing of the MH hypothesis with empirical evidence. Although GS appears to correlate negatively with population size in eukaryotes (Lynch and Conery 2003), in agreement with the MH hypothesis, this relationship vanishes when accounting for phylogenetic non-independence among taxa (Whitney and Garland 2010). In addition, predictions of the MH hypothesis can vary in opposite directions depending on the underlying pattern of indel mutations. In eukaryotes, where indel mutation patterns are typically biased toward insertions, reductions in population size are predicted to lead to increasing GS. Conversely, similar N_e reductions are expected to result in decreasing GS in bacteria, for which the mutation pattern is biased toward deletions (Kuo et al. 2009). Moreover, although essential to the MH hypothesis, N_e remains difficult to estimate. Most studies rely on population polymorphism (Lynch and Conery 2003) or heterozygosity (Yi and Streelman 2005), two measures that typically reflect population size history over the last tens of thousands to millions of generations, whereas the pace of genome evolution might take place at much longer temporal scales (Whitney and Garland 2010; Whitney et al. 2011).

Since the formulation of the MH hypothesis, the few early empirical studies that originally supported a relationship between GS and N_e (Lynch and Conery 2003; Yi and Streelman 2005) have been criticized (Gregory and Witt 2008; Whitney and Garland 2010), and later analyses failed to support this relationship. No relationships were found between GS and (1) allozyme polymorphism in plants (Whitney et al. 2010), (2) molecular polymorphism among different species of rice (genus *Oryza*) (Ai et al. 2012), (3) the relative population size in seed beetles (Arnqvist et al. 2015), and (4) genetic drift in salamanders compared to frogs (Mohlhenrich and Mueller 2016). Finally, the influence of N_e on *Caenorhabditis* has been dismissed in favor of an adaptive explanation (Fierst et al. 2015). However, all these studies suffered either from the use of very indirect proxies of N_e or from small gene samples, often characterized for not more than 12 species (although exceptions exist, see Whitney et al. 2010). Therefore, the influence of N_e on GS remains to be tested on an empirical data set that provides a robust estimate of N_e within a statistically powerful framework.

Habitat shifts often result in drastic changes in population size and therefore offer useful case studies for testing the MH hypothesis. In this study, we use a comparative genomic approach to test whether nonadaptive forces drive changes in GS following the habitat shift from surface water to groundwater within asellid isopods. The colonization of groundwater from surface water took place at multiple times and locations over the last tens to hundreds of million years within this family (Morvan et al. 2013), thereby providing independent replicates of the transition to dark and low-energy habitats (Huntsman et al. 2011; Venarsky et al. 2014). Groundwater colonization leads to eye-degeneration and is considered irreversible (Niemiller et al. 2013). We use 11 pairs of closely related surface and subterranean asellid species to test the predictions of the MH hypothesis (Supplemental Table S1). According to the MH hypothesis and assuming that consistent population size reduction took place following groundwater colonization, then subterranean species are predicted to show reduced selection efficacy and larger GS than their surface relatives. We also considered alternative hypotheses, namely (1) the possible reduction in GS in response to energy limitation in groundwater, and (2) the selection of particular life history traits (hereafter, growth rate and body size) as a driver of patterns of GS variation.

Results

Efficacy of natural selection in groundwater

To evaluate differences in selection efficacy between surface and subterranean species, we sequenced and de novo assembled the transcriptomes of 11 pairs of asellid species. After gene family delimitation, we estimated the rate of nonsynonymous over synonymous substitutions (d_N/d_S) on a set of conserved and single-copy genes. This ratio is jointly defined by the distribution of selection coefficient of new mutations (s) and the magnitude of genetic drift as defined by N_e (Nielsen and Yang 2003). Therefore, the transcriptome-wide d_N/d_S is expected to increase over extended periods of small N_e because of the increasing fixation of slightly deleterious mutations (Ohta 1992), an expectation confirmed in a wide range of animals (Galtier 2016). Consequently, the transcriptome-wide d_N/d_S is a direct proxy of selection efficacy. Subterranean species show significantly higher transcriptome-wide d_N/d_S than their surface relatives (Table 1; Fig. 1). Looking at each pair of species independently, eight of 11 pairs display a higher subterranean transcriptome-wide d_N/d_S, a relative increase that can be as high as 59% (Fig. 1).

Although long periods of reduced N_e will induce higher transcriptome-wide d_N/d_S , adaptation to new habitats could potentially produce the same effect. Under the action of positive selection, beneficial nonsynonymous mutations will reach fixation faster than their synonymous counterparts and will lead to sites with $d_N/d_S > 1$. If the frequency of such sites increases during the transition to groundwater, then we can expect the transcriptome-wide d_N/d_S to increase. We first tested this adaptive hypothesis using a model that allows d_N/d_S variation across sites and makes it possible to differentiate between variation in the intensity of purifying (w^-) and positive selection (w^+) and their respective frequencies. Subterranean species do not show an elevated frequency [$fq(w^+)$] or intensity of positive selection (w^+) but show higher w^- (Table 1; Supplemental Table S2; Supplemental Figs. S1, S3). This supports a scenario in which subterranean species do not experience higher rates of positive selection, but instead evolve under reduced purifying selection efficacy.

We next tested the adaptive d_N/d_S increase scenario using polymorphism data. Under a high rate of positive selection with recurrent fixation of nonsynonymous mutations, populations will display an excess of nonsynonymous substitutions compared to nonsynonymous polymorphism (McDonald and Kreitman 1991). We used the “direction of selection” statistics (DoS) (Stoletzki and Eyre-Walker 2011), which is a transcriptome-wide comparison of the rates of nonsynonymous substitution and polymorphism, to test if positive selection indeed led to higher rates of fixation in subterranean species ($DoS > 0$). Most subterranean species have negative *DoS* (Supplemental Fig. S4), and subterranean species do not have higher *DoS* than surface species (Table 1). On the contrary, in most species, irrespective of their habitat, the *DoS* is close to 0 or negative, indicating that many slightly deleterious mutations are segregating in these species. Although this observation is in line with the idea that effective population size reduces the efficacy of selection in these species, it does not completely rule out the hypothesis that subterranean species may have concomitantly evolved higher d_N/d_S as a result of more frequent adaptations during the shift from surface to subterranean habitats. When slightly deleterious mutations dominate the evolutionary dynamics, which appears to be the case in this group, they can mask the influence of adaptive evolution on polymorphism (James et al. 2016). We further tested this hypothesis

Table 1. Phylogenetic generalized least-squares (PGLS) models testing the correlation between two variables

Dependent variable	Predictor variable	*n*	LRT *P*-value	Coefficient	AIC	R^2
d_N/d_S	Ecological status	22	0.015[a]	0.017		0.237
w^-		22	0.020[a]	0.018		0.217
w^+		22	0.355	−0.197		0.038
$fq(w^+)$		22	0.653	0.001		0.009
DoS		22	0.410	0.056		0.030
$\hat{\theta}_w$		22	0.099	−0.001		0.117
p_N/p_S		22	0.232	0.036		0.063
Growth rate		16	0.011[a]	−6.211		0.332
Body size		22	0.192	−1.135		0.074
d_N/d_S	Relative colonization time	19	0.001[b]	0.064		0.427
w^-		19	0.001[b]	0.073		0.420
w^+		19	0.521	−0.465		0.021
$fq(w^+)$	Colonization time	19	0.016[a]	0.006		0.263
DoS		19	0.817	0.035		0.003
$\hat{\theta}_w$		19	0.486	−0.001		0.025
p_N/p_S		19	0.355	0.058		0.044
Growth rate		13	0.115	−7.813		0.174
Body size		19	0.581	−1.134		0.016
Genome size	Ecological status	22	0.014[a]	0.340	29.1	0.240
	Colonization time	19	0.019[a]	0.789		0.250
	d_N/d_S	22	0.009[b]	10.447	28.3	0.266
	w^-	22	0.004[b]	9.897	26.8	0.315
	$\hat{\theta}_w$	22	0.450	−64.920	34.6	0.026
	p_N/p_S	22	0.428	0.823	37.8	0.028
	Growth rate	16	0.178	−0.022		0.107
	Body size	22	0.862	0.006	35.1	0.001
Repeatome size	Ecological status	22	0.017[a]	0.251	29.1	0.229
	Colonization time	19	0.003[b]	0.724		0.372
	d_N/d_S	22	0.006[b]	8.189	15.3	0.287
	w^-	22	0.002[b]	7.808	13.5	0.344
	$\hat{\theta}_w$	22	0.437	−50.457	22.2	0.027
	p_N/p_S	22	0.419	0.633	22.1	0.029
	Growth rate	16	0.148	−0.017		0.123
	Body size	22	0.916	0.003	22.8	0.001

On the *top* of the table, correlation tests between the transition to groundwater (ecological status, proportion of subterranean branch, or colonization time) and variables ranging from selection efficacy (d_N/d_S, w^-), rate of adaptive evolution [w^+, $fq(w^+)$, *DoS*], polymorphism ($\hat{\theta}_w$ and p_N/p_S), and phenotypic traits (growth rate and body size) are reported. Another set of correlation tests between GS, or repeatome size, and some of these variables is also reported at the *bottom* of the table. Coefficients are in contrast to the surface status. Only comparable AIC are shown (same dependent variable and same number of observations).

(w^-) intensity of purifying selection; (w^+) intensity of positive selection; [$fq(w^+)$] frequency of sites under positive selection; (*DoS*) direction of selection; (LRT *P*-value) likelihood ratio test between the models with and without the given predictor variable; (R^2) Cox and Snell generalized R^2; (*n*) number of observations; (relative colonization time) represents the proportion of the terminal branches estimated to be subterranean using the opsin gene, it is equal to $\text{time}_{\text{colonization}}/\text{time}_{\text{speciation}}$.

[a] *P*-value <0.05.

[b] *P*-value <0.01.

by directly estimating the rate of adaptation (α) in two species pairs where subterranean species display elevated d_N/d_S compared to their surface sister species (*P. beticus* versus *P. jalionacus* and *P. coiffaiti* versus *P. cavaticus*) (Fig. 1). We used a McDonald-Kreitman modified approach (McDonald and Kreitman 1991; Messer and Petrov 2013) designed to cancel the influence of demography and linkage effects. For species such as those studied here, the high prevalence of segregating deleterious mutations at low frequency inflates the rate of nonsynonymous polymorphism and artificially decreases α estimates. Messer and Petrov (2013) suggest to reconstruct α as a function of the derived allele frequencies. As allele frequency increases, slightly deleterious alleles become rarer, to the point that a robust estimate of α can be obtained by calculating the asymptotic α for an allele frequency of 1. By resequencing the transcriptomes of 4–5 individuals per species, we reconstructed the unfolded site frequencies of these two pairs of species, fitted the distribution of α(*x*), and estimated the asymptotic α(1). For each species, we recovered the expected distribution of α(*x*) as in Messer and Petrov (2013) and obtained asymptotic α that were negative (Supplemental Fig. S5). Although the subterranean species of these two pairs of species show clear transcriptome-wide d_N/d_S increases, they do not display elevated rates of adaptation. For one pair, there is no significant α variation (*P. beticus–P. jalionacus*; *P*-value = 0.49), whereas in the other pair the subterranean species shows lower α estimates (*P. coiffaiti–P. cavaticus*; *P*-value = 0.02). Therefore both *DoS* and α analyses confirm that the increase in d_N/d_S in subterranean species is not caused by a higher rate of positive selection, in line with the results found on the model differentiating between variation in the intensity of purifying (w^-) and positive selection (w^+).

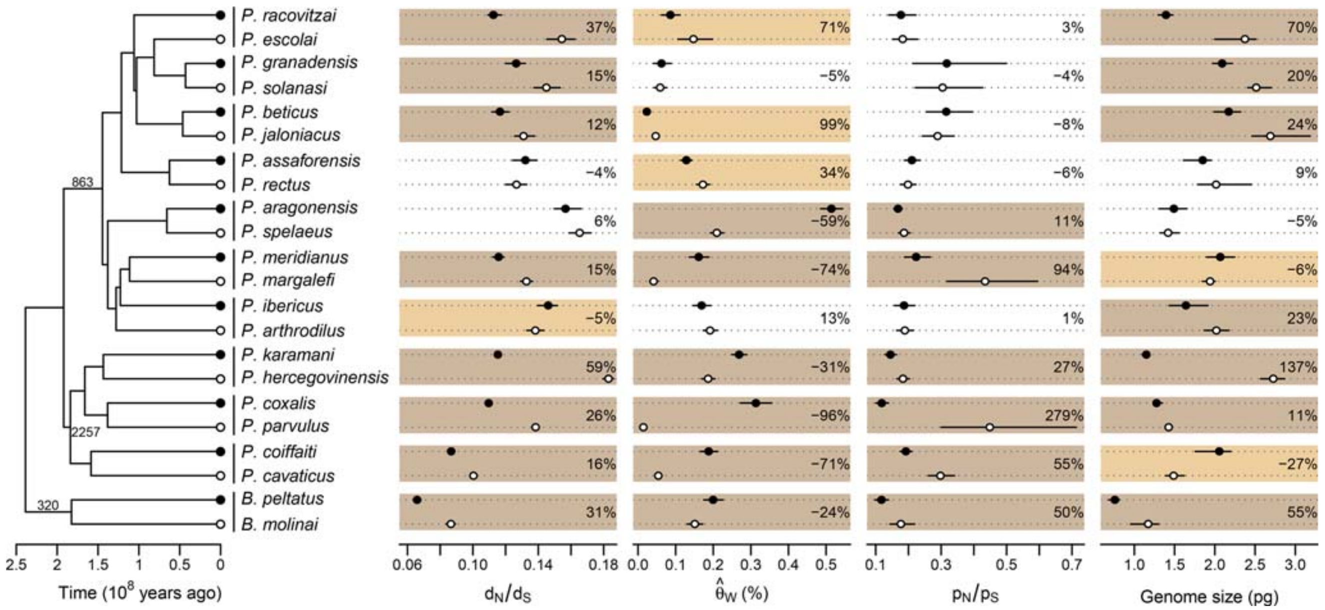

Figure 1. Selection efficacy (d_N/d_S), polymorphism ($\hat{\theta}_w$ and p_N/p_S), and haploid genome size measurements for 11 pairs of surface and subterranean asellid species. Vertical bars next to the tree indicate species pairs with their surface (black circles) and subterranean (white circles) species. Numbers along branches of the tree are the numbers of single-copy genes used to estimate the d_N/d_S. Color boxes indicate statistical support (P-value <0.05) in favor of (dark brown) or against (light brown) a decrease in selection efficacy or population size, or an increase in genome size in subterranean species. No box indicates no statistical differences between species of a pair. Error bars represent 95% bootstrap confidence intervals, except for genome size where it represents the range around the mean for five individuals. The percentage change from surface water to subterranean species is shown for each species pair.

Subterranean species share multiple convergent regressive phenotypes, such as the loss of eyes and pigmentation that may ultimately be associated with gene nonfunctionalizations (Protas et al. 2005; Niemiller et al. 2013). A transcriptome-wide d_N/d_S increase can therefore also be caused by an excess of genes that have lost their function and consequently have acquired d_N/d_S nearing 1. As an example, the opsin gene of subterranean species has a much higher d_N/d_S as a result of gene nonfunctionalization (see next section and Supplemental Fig. S6). Release of functional constraint on a gene and the resulting d_N/d_S increase should also be paralleled with much lower gene expression, or no expression at all (Zou et al. 2009; Yang et al. 2011). This is typically observed for the opsin gene, which has much lower expression in the subterranean species (Supplemental Fig. S6). If gene nonfunctionalization in subterranean species is pervasive enough to shift the transcriptome-wide d_N/d_S upward, we expect to see a subset of genes with lower expression in the subterranean species. We tested this hypothesis by comparing the expression of the genes in the surface and subterranean species of each pair. The set of conserved and single-copy genes that was used to calculate each species d_N/d_S has much higher expression levels than the complete transcriptomes (Supplemental Fig. S7). This set of genes also displays more conserved gene expression levels across species (Supplemental Fig. S7). Finally, after counting changes in gene expression category between sister species, we found no evidence of an excess of genes with lower expression in the subterranean species (Wilcoxon signed rank test P-value = 0.650) (Supplemental Table S4). We further tested this nonfunctionalization hypothesis by looking for a subset of genes that display larger d_N/d_S in the subterranean species, while the remaining genes display no d_N/d_S variation. Distributions of the variation in gene d_N/d_S did not support the existence of such a small subset of genes (Supplemental Fig. S2). On the contrary, these distributions were unimodal with a median positively correlated to the transcriptome-wide d_N/d_S ($R^2 = 0.62$, P-value = 0.004).

Altogether, we accumulated multiple evidences that the transcriptome-wide d_N/d_S increase observed in subterranean species is not the consequence of increased levels of positive selection or gene nonfunctionalization, but rather the result of convergent reductions in the efficacy of purifying selection among subterranean species.

Polymorphism proxies of N_e

Instead of directly assessing selection efficacy, the MH hypothesis has traditionally been tested using polymorphism data. Indeed, polymorphisms provide a direct proxy for N_e, which tunes the magnitude of genetic drift and ultimately the efficacy of selection. As transcriptomes were sequenced from pooled individuals, we estimated synonymous and nonsynonymous polymorphism for genes with high coverage. We used the population mutation rate ($\hat{\theta}_w$) which is proportional to the product of N_e and the mutation rate μ, and the ratio of nonsynonymous over synonymous polymorphism (p_N/p_S), which is expected to decrease with increasing N_e, independently of μ. Both the d_N/d_S and the p_N/p_S measure the efficacy of selection to purge slightly deleterious mutations, although the latter works at a much shorter timescale. As expected, $\hat{\theta}_w$ and p_N/p_S are negatively correlated (phylogenetic generalized least-squares [PGLS] models, P-value <0.001, $R^2 = 0.43$). Polymorphism data are generally consistent with selection patterns inferred from d_N/d_S: subterranean species have significantly higher p_N/p_S than their surface relatives in six of 11 pairs, whereas there is no pair significantly supporting the opposite pattern (Fig. 1).

However for $\hat{\theta}_w$ the pattern is less clear: in six pairs, subterranean species have significantly lower $\hat{\theta}_w$, whereas in three pairs, subterranean species show significantly higher $\hat{\theta}_w$ (Fig. 1). Overall, the differences in p_N/p_S or $\hat{\theta}_w$ between subterranean and surface species are not statistically significant (Table 1). In addition, there is no correlation between $\hat{\theta}_w$ or p_N/p_S and the efficacy of selection as estimated using the transcriptome-wide d_N/d_S (PGLS P-value = 0.80 and 0.79, respectively). Traditional polymorphism proxies of N_e do not therefore support the same scenario as the one depicted using selection efficacy (d_N/d_S).

Estimating colonization times with opsin sequences

Subterranean species may have colonized groundwater at different time periods, some being subterranean for a much longer time than others. Ignoring such differences through the use of a qualitative present-day ecological status (i.e., surface versus subterranean) may limit our power to detect a change in GS or polymorphism associated with the subterranean transition. One could contrast polymorphism measures and the time since the latest speciation event, where we know that a species ancestor was a surface species, but this would only be valid if speciation and colonization times were synchronous. Alternatively, we estimated the colonization time using the nonfunctionalization of the opsin gene. Indeed, similarly to observations made in underground mammals (Emerling and Springer 2014), together with the regression of the ocular system, some subterranean species display loss-of-function mutations in eye pigment (Leys et al. 2005) or opsin genes (Niemiller et al. 2013), which are indicative of a loss of functional constraint. If we assume that opsin gene sequences have lost their function early in the process of groundwater colonization, then they must have been evolving under a neutral model ($d_N/d_S = 1$) since that colonization. Using a two-states model of evolution, with one surface opsin d_N/d_S estimated using opsins from surface species, and one subterranean opsin d_N/d_S equal to 1, we can then estimate the colonization time as a function of the speciation time and the estimated opsin d_N/d_S measured on a given branch leading to a subterranean species.

Using a combination of Sanger sequencing, transcriptome assemblies, and genome sequencing reads, we reconstructed one opsin ortholog for 19 of 22 species. Irrespective of their ecological status, the two species of the genus *Bragasellus* probably do not possess this opsin locus. In addition, for one *Proasellus* subterranean species (*P. parvulus*), failure to amplify or recover Illumina reads from this locus suggests that the whole locus was lost in this species. Subterranean species showed lower opsin expression levels and had much higher opsin d_N/d_S ratios than surface species (average $d_N/d_S = 0.3$ and 0.05, respectively) (Supplemental Fig. S6). In addition to one subterranean species, which completely lost the locus (*P. parvulus*), two subterranean species also harbored clear nonfunctionalization signatures consisting of an 18-base-long deletion for *P. solanasi* and the insertion of a 280-base-long repeat element in the sequence of *P. cavaticus*. These observations validate the opsin locus as a colonization clock.

Estimated colonization times vary greatly, with more than 50× variation between the youngest subterranean species (*P. jalionacus*, 2 MYA) and the oldest one (*P. herzegovinensis*, 122 MYA) (Supplemental Table S5). Colonization time is related to the regression of the eye and pigmentation, with species with intermediate phenotypes (reduced eyes and partial depigmentation) being very recent subterranean species (Supplemental Fig. S8). Using relative colonization time ($\text{time}_{\text{colonization}}/\text{time}_{\text{speciation}}$) for d_N/d_S ratios or absolute colonization time instead of the present-day ecological status gives very similar results (Table 1), indicating that variation in the colonization time is not likely to obfuscate polymorphism variation. Conversely, the strength of the correlation between the transcriptome-wide d_N/d_S (or w^-) and relative colonization time is higher than with the ecological status (Table 1), reinforcing the hypothesis of a causal link between the subterranean colonization and the subsequent drop in selection efficacy. The only exception is the frequency of sites under positive selection [$fq(w^+)$] (Table 1), which becomes significantly higher in subterranean species when colonization time is used instead of ecological status (PGLS P-value = 0.016, +0.6% per 100 million years of colonization).

Genome size increase in groundwater

We next measured genome sizes in all 11 species pairs using flow cytometry. Using either the ecological status or the colonization time, we found a statistically significant increase in GS following the transition from surface to groundwater habitats (Table 1; Fig. 1). Looking at each pair of species independently, seven of 11 pairs display a significantly higher GS, a relative increase reaching 137% in *P. hercegovinensis* (Fig. 1). This finding is robust to the addition of 19 asellid species (PGLS with 41 asellids, P-value = 0.022, coefficient = 0.273) and to the inclusion of a wider range of metazoans (linear mixed model with 18 independent pairs, including Decapoda and Gastropoda; P-value = 0.040, 25% average increase in GS) (Fig. 2; Supplemental Table S6).

Linking genome size to selection efficacy

One of the main predictions of the MH hypothesis is that GS is negatively correlated with selection efficacy in eukaryotes. We validated this prediction because we found a highly significant positive relationship between GS and the transcriptome-wide d_N/d_S (or w^-) (Table 1; Supplemental Fig. S9). In addition, the d_N/d_S ratio (or w^-) achieves similar, if not better, performance in predicting GS variation than the ecological status or colonization time (lower AIC and higher R^2) (Table 1). When d_N/d_S (or w^-) is put first and ecological status second into a single PGLS model of GS, the effect of the ecological status is no longer significant (PGLS P-value = 0.189).

Testing other covariates

In contradiction to the MH hypothesis, adaptative hypotheses postulate that variation in GS is under direct selection via its impacts on cellular (such as nucleus and cell sizes) and organismal parameters (such as body size and growth rate) (Gregory 2001). In many species, population size covaries with traits under selection such as growth rate and body size, themselves correlated to some extent to GS, making any causation test extremely challenging (Gregory 2005). Although body size was readily available in the literature, we estimated growth rate in 16 species using the RNA/protein ratio, which is known to be positively correlated to growth rate in Rotifera (Wojewodzic et al. 2011). Indeed, in situ estimates of growth rates were out of reach, and a more traditional proxy such as the RNA/DNA ratio is inapplicable when GS varies. In accordance to the general assumption that subterranean animals tend to adopt K-selection life history traits, subterranean asellids species display lower growth rate, although no trend was found regarding body size (Table 1). However, growth rate and body size do not correlate with GS (Table 1; Supplemental Fig. S9). Thus, although many forces might be at play during the transition to

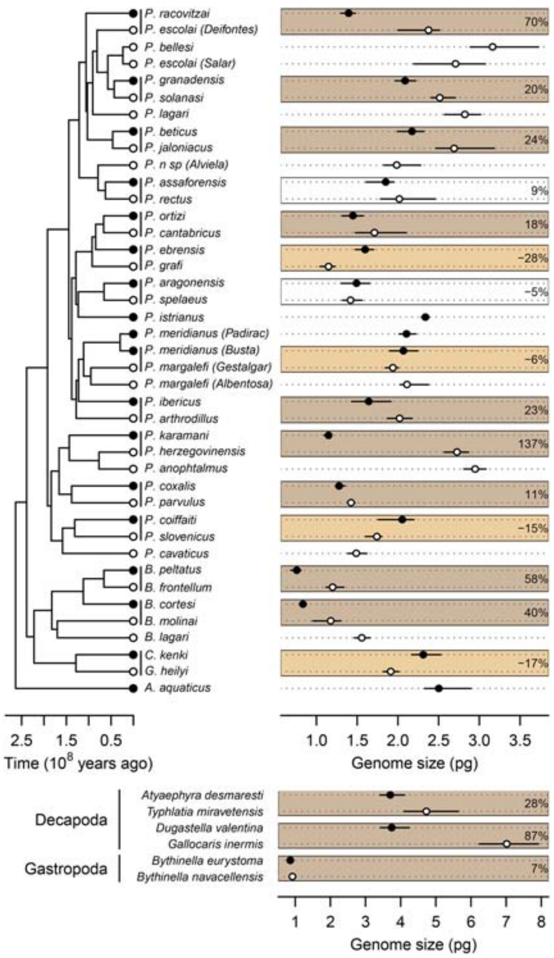

Figure 2. Variation in haploid genome size associated with the ecological transition from surface water to groundwater in 47 species, including isopods (*top*) and decapods and gastropods (*bottom*). The 18 independent pairs of surface and subterranean species are delimited with boxes. Legends as in Figure 1. Identical species names followed by locality names within brackets refer to cryptic species (Morvan et al. 2013).

groundwater habitats, in asellids, we only found correlation between selection efficacy and GS.

Mechanism of genome size increase

Implicit in the MH hypothesis is that an increase in GS should result from the progressive spread of insertions with slightly deleterious fitness effects, such as transposable elements (Vieira et al. 2002). Yet, other much faster mechanisms such as polyploidization events can also inflate GS (Otto 2007). We tested for the occurrence of such large duplication events by looking for an excess of recent paralogs in the 11 subterranean species compared to their surface sister species. The mean number of gene copies per gene family is not correlated to the ecological status nor to GS (PGLS P-value = 0.773 and 0.579, respectively) (Supplemental Fig. S10), indicating that subterranean species do not present an excess of recent duplication events.

To test for the accumulation of repeat elements, we evaluated the amount of repetitive DNA in the 11 asellid species pairs using low-coverage genome sequencing, followed by clustering of highly repetitive elements. Indeed, contrary to the non-repetitive fraction of the genome, elements at high frequency will collect enough reads to be assembled. Summing across the contributions of each element provides an estimate for the size of the repeatome (i.e., the fraction of the genome consisting of repeated DNA elements). We found larger repeatomes in large genomes (Table 2; Fig. 3A,D; Supplemental Fig. S11). The repeatomes are largely made of repeat families found in a single species, called repeat orphans, with very few shared repeats across species (Fig. 3B). The occurrence of these shared repeats is largely explained by phylogenetic relatedness: closely related sister species share more than 200 repeat families, with this number quickly decreasing with divergence time (Fig. 3C). GS has little power to explain the composition of the repeatome. None of the axes of a repeatome composition correspondence analysis are correlated to GS, whereas the first three axes harbor a strong phylogenetic signal (Blomberg K > 1 with P-value <0.01) (Supplemental Table S7; Supplemental Fig. S12).

The pattern of GS increase is globally congruent with a global increase of the repeatome invasiveness. Indeed, big genomes have at the same time more repeats and repeats at higher frequencies (Table 2; Fig. 3E,D). To a lesser extent, the number of repeat orphans and their frequencies also increases with GS (Table 2), demonstrating that big genomes are also more prone to genome invasion by new repeats. Nonetheless, the ratio of the total genomic size (TGS) occupied by new repeats over common repeats does not change ($\mathrm{TGS}_{\mathrm{orphans}}/\mathrm{TGS}_{\mathrm{non-orphans}}$) (Table 2), indicating that this aspect of the repeat community structure does not change as GS increases. So, in contrast to several model organisms such as humans or maize, the GS increase was not induced by a very limited set of elements, but is the consequence of a repeat element community that became globally more invasive subsequent to the ecological transition.

Although the repetitive portion of the genome increases linearly with GS, it does not explain 100% of GS variation: on average 1 Gb of repeats was gained for 1.3 Gb of GS increase (Table 2). Consequently, the estimated TGS of the non-repetitive portion of the genome also increases with GS, though at a much slower pace (1 Gb for 2.8 Gb) (Table 2). Either repeats are harder to assemble in large genomes, or another minor mechanism is also at play during GS increase. Directly using the repeatome size instead of GS in correlation analyses gives similar or reinforced results: although polymorphism-based N_e proxies ($\hat{\theta}_w$ or p_N/p_S), growth rate, and body size do not correlate with repeatome size, selection efficacy (d_N/d_S or w^-) does (Table 1).

Discussion

We found a substantial correlation between selection efficacy, as measured by transcriptome-wide d_N/d_S, and repeatome size. This finding indicates that, for a large part, GS is controlled by the efficacy of selection to prevent the invasion of the genome by repeat elements. Conversely, we found no correlation between N_e estimates derived from polymorphism data and GS. At first glance, this result sounds contradictory since the efficacy of selection depends on N_e. We propose two non-mutually exclusive hypotheses to explain this contradiction. First, although the transcriptome-wide d_N/d_S provides an average estimate of selection efficacy since the divergence of two species of a pair, polymorphism-based proxies such as $\hat{\theta}_w$ or p_N/p_S are influenced by recent N_e fluctuations, independently of the divergence time. If N_e fluctuates rapidly with large amplitude around a stable mean, polymorphism is likely to provide a noisy proxy of this mean, contrary to the d_N/d_S. This

Table 2. Phylogenetic generalized least-squares models testing the association between genome size (dependent variable) and the size and composition of the 22 species repeatomes

Predictor variable	LRT *P*-value	Coefficient	AIC	R^2
Repeatome size (Gb)	<0.0001[a]	1.295	−33.0	0.955
Non-repetitive genome size (Gb)	<0.0001[a]	2.773	6.5	0.728
Number of repeats	<0.0001[a]	0.007	3.7	0.760
top1 TGS (Gb)	0.0022[a]	20.546	25.7	0.348
top10 TGS (Gb)	<0.0001[a]	5.407	1.6	0.783
Repeat orphans TGS (Gb)	0.0163[b]	0.002	29.4	0.231
Number of orphan repeats	0.0056[a]	0.003	27.5	0.294
Orphans TGS/nonorphans TGS	0.3797	−0.203	34.4	0.034

Same abbreviations as in Table 1.
(TGS) total genomic size; (Gb) gigabase; (top1) most invasive repeat; (orphan) repeat element found in one species only.
[a] *P*-value <0.01.
[b] *P*-value <0.05.

hypothesis is supported by larger coefficients of variation for $\hat{\theta}_w$ or p_N/p_S than for the d_N/d_S ($CV_{d_N/d_S} = 0.22$, $CV_{p_N/p_S} = 0.39$, $CV_{\hat{\theta}_w} = 0.73$, test of the equality of CVs *P*-value <0.001). In this study, we observed the effect of multiple groundwater transitions that happened tens to hundreds of million years ago, a time scale long enough to produce large d_N/d_S and GS variations, but also encompassing important climatic fluctuations, which potentially generated shorter time-scale N_e variations. Particularly, quaternary climatic fluctuations are likely to have produced important N_e fluctuations in surface species, which have much more unstable habitats than subterranean species. Therefore, the lack of a clear correlation between short-term N_e proxies, like polymorphism, and GS might be the consequence of recent climatic fluctuations. Interestingly, this hypothesis received some support using Tajima's *D* tests. Indeed, only *P. beticus* (a surface species) of the four species for which we have adequate data to estimate Tajima's *D* is not at the mutation-drift equilibrium (*P*-value = 0.017) (Supplemental Table S8). This surface species shows evidence of recent population contraction (Tajima's $D > 0$), which might explain its unexpected combination of low polymorphism and low d_N/d_S when compared to its sister subterranean species (Fig. 1). Second, polymorphism variation might also be more prone to measurement artifacts than d_N/d_S. In particular, SNP calling errors can constitute a relatively important fraction of detected variants among species with low levels of polymorphism. For most species pairs (7 of 11), we observed a higher level of polymorphism in surface than in subterranean species (Fig. 1; Supplemental Fig. S13). The four other cases all correspond to pairs for which both species have low level of polymorphism, which might therefore be subject to higher measurement error rates. This is in line with the strong relationship observed between the surface species $\hat{\theta}_w$ and the difference in $\hat{\theta}_w$ between species pairs ($R^2 = 0.74$, *P*-value <0.001) (Supplemental Fig. S13). This suggests that below

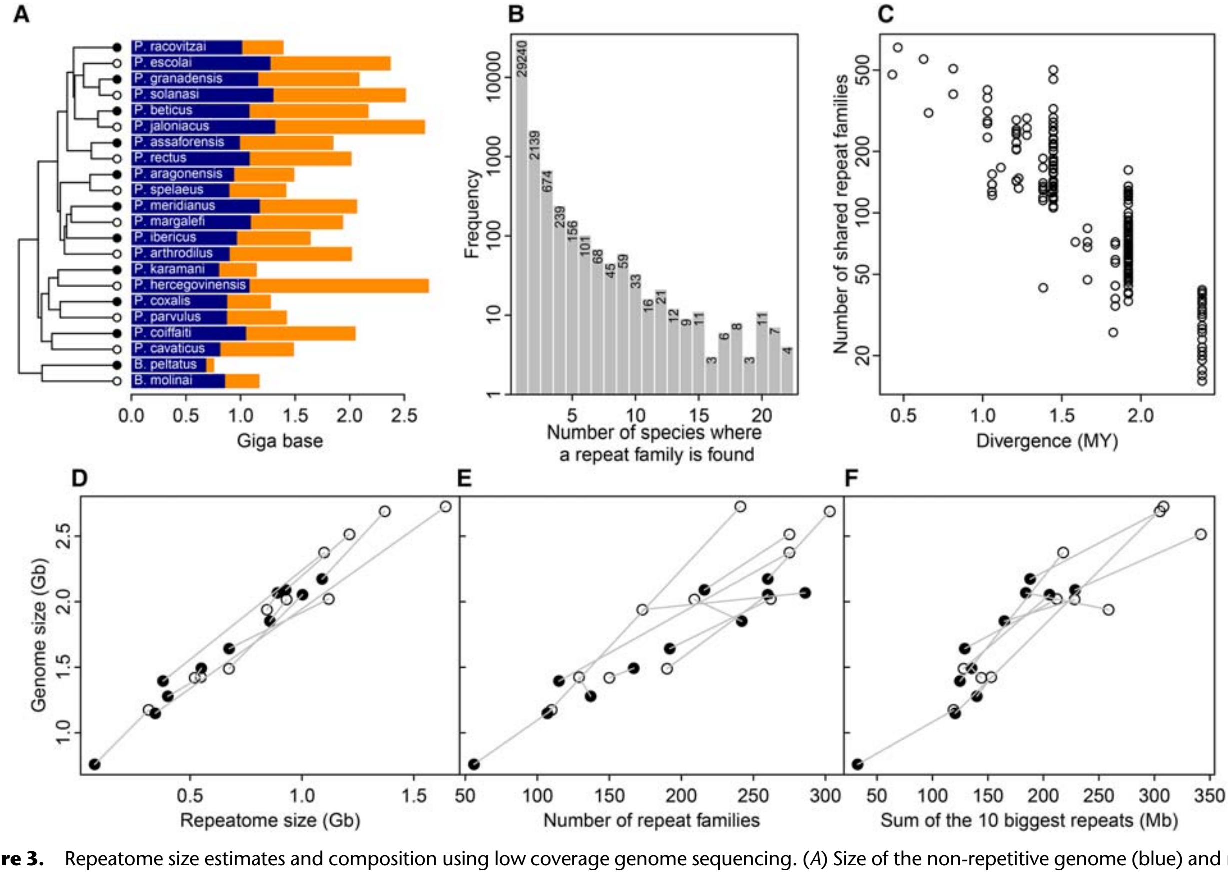

Figure 3. Repeatome size estimates and composition using low coverage genome sequencing. (*A*) Size of the non-repetitive genome (blue) and repeatome (orange) for the 22 species (tree symbols as in Figure 1). (*B*) Repeat family frequency spectrum. (*C*) Number of shared repeats between two species as a function of divergence time. Relationship between GS and repeatome size (*D*), the number of repeat families (*E*), and the total genomic size of the 10 biggest repeat families (*F*). In *D*, *E*, and *F*, surface (black circles) and subterranean (white circles) species of a pair are joined by a gray line.

approximately 2, $\hat{\theta}_W$ becomes a poor indicator of N_e changes. Altogether, the use of polymorphism proxies of N_e for polymorphism-poor species or for species that experienced recent N_e variations might therefore result in misleading rejection of the MH hypothesis.

Our findings shed new light on the debate of the validity of the MH hypothesis and the comparative methods that should be implemented to test it (Charlesworth and Barton 2004; Gregory and Witt 2008; Whitney and Garland 2010; Lynch 2011; Whitney et al. 2011). Using a relatively reduced set of ecologically contrasted species pairs as true replicates of the same ecological transition is statistically more powerful than testing for differences in genomic attributes among a larger set of distantly related taxa, in which the number of independent observations is unknown and for which many traits varies. Accounting for phylogenetic effects in statistical analyses of GS variation among multiple species is another yet crucial aspect because it increases not only specificity (Whitney and Garland 2010) but also sensitivity. Taken all together, subterranean species do not have larger GS than surface species (ordinary least square linear model P-value = 0.122 for the 11 species pairs, P-value = 0.095 for the 41 asellid species, P-value = 0.261 for the 18 metazoan species pairs), although pairwise comparisons of surface and subterranean species (Figs. 1, 2) and PGLS models (Table 1) reveal a very clear and significant pattern of higher GS among subterranean species.

Despite several lines of evidence supporting a lower selection efficacy caused by long-term N_e reduction in subterranean species, we found little support that the colonization of this new habitat was also paralleled with adaptive evolution. The only evidence was found in the frequency of sites under positive selection when colonization time was used. The increase was nonetheless moderate (+0.6% per 100 MY of colonization) and was not supported by polymorphism (*DoS* or α) analyses. However, this study is limited to a small set of gene families that are found in most species, in a single copy, and whose expression is very conserved. This set of genes is probably under strong purifying selection and might be less prone to positive selection. Fully investigating the relative role of adaptive versus nonadaptive forces during this ecological transition will require a much broader genomic approach.

Disentangling the forces that drive GS variation has commonly been complicated by rampant covariation between GS and multiple traits such as cell and body sizes, growth rates, metabolism and N_e, to name a few. In this study, we found no association between GS and two common covariates (body size and growth rate). Although we cannot completely rule out other nontested parameters and alternative ad hoc adaptive hypotheses, the results of this study are fully compatible and best explained by a causal relationship between N_e and GS. The mechanisms that drive genome size variation are also fully compatible with the MH hypotheses. The repeat elements were globally more diversified and more invasive in species with reduced selection efficacy, an expected outcome if selection against repeat element proliferation is less effective.

Documenting changes in the architecture of genomes among taxa that have undergone major shifts in habitats (Protas et al. 2005; Jones et al. 2012; Fang et al. 2014; Soria-Carrasco et al. 2014) holds much promise for disentangling evolutionary processes driving genome evolution. In accordance with the MH hypothesis, our focus on the genomics of groundwater colonization brings new evidence for a prominent role of nonadaptive forces in GS evolution. Despite strong energetic constraints in groundwater, GS likely increases under the long-term effect of reduced N_e, which limits the strength of natural selection in hampering the invasion of slightly deleterious repeat elements. Altogether this study supports long-term effective population size variation as a key evolutionary regulator of genome features.

Methods

Aselloidea timetree

Phylogenetic comparative methods require accurate estimates of phylogenetic relationships and divergence times among species (Purvis et al. 1994). Both were inferred from a large timetree of Aselloidea containing 193 evolutionary units (Morvan et al. 2013; Supplemental Table S9). Sequences of the mitochondrial cytochrome oxidase subunit I (*COI*) gene, the 16S mitochondrial rDNA gene, and the 28S nuclear rDNA gene used to build the Aselloidea timetree were obtained according to previously described methods (Calvignac et al. 2011; Morvan et al. 2013). Alignments and Bayesian estimates of divergence times were conducted according to Morvan et al. (2013). From the Aselloidea timetree, we selected 11 independent pairs of surface and subterranean asellid species as replicates of the ecological transition from surface water to groundwater.

RNA-seq

For the 11 selected species pairs (Supplemental Table S1), individuals were sampled from caves, springs, wells, and the hyporheic zone of streams using different pumping and filtering devices (Bou-Rouch pump, Cvetkov net, and Surber sampler) and were flash frozen alive. Total RNA was isolated using TRI Reagent (Molecular Research Center). Extraction quality was checked on a Bioanalyser RNA chip (Agilent Technologies), and RNA concentrations were estimated using a Qubit fluorometer (Thermo Fisher Scientific). Prior to any additional analysis, species identification was corroborated for each individual by sequencing a fragment of 16S gene. Equimolar pools of at least five individuals were made to achieve 10 µg of the total RNA (Supplemental Table S1). Volumes were reduced using a Concentrator-Plus (Eppendorf) to achieve approximately 10 µL. Double-strand poly(A)-enriched cDNA were then produced using the Mint2 kit (Evrogen) following the manufacturer protocol except for the first-strand cDNA synthesis, in which the CDS-1 adapter was used with the plugOligo-Adapter of the Mint1 kit (5′-AAGCAGTGGTATCAACGCAGAGTACGGGGG_p-3′). After sonication with a Bioruptor Nextgen UCD300 (Diagenode) and purification with MinElute (Qiagen), Illumina libraries were prepared using the NEBNext kit (New England BioLabs) and amplified using 22 unique indexed primers. After purification with MinElute, 400–500 bp fragments were size selected on an agarose gel. Libraries were paired-end sequenced on a HiSeq 2000 sequencer (Illumina) using 100 cycles at the Danish National High-throughput DNA Sequencing Center (Copenhagen, Denmark). A full lane was used for one species (*Proasellus beticus*), and reads were resampled to represent ∼2%, 5%, 10%, 25%, 50%, and 100% of a full lane. These six sets of reads were de novo assembled (see next section) and the number of assembled components >1 kb was compared among sets (Supplemental Fig. S14). This preliminary experiment was used as a rational procedure to multiplex four species on one lane.

Transcriptome assembly

Adapters were clipped from the sequence, low-quality read ends were trimmed (phred score <30), and low-quality reads were discarded (mean phred score <25 or if remaining length <19 bp) using fastq-mcf of the ea-utils package (Aronesty 2013). Transcriptomes were de novo assembled using Trinity (version 2013-02-25) (Grabherr et al. 2011). Open reading frames (ORF) were identified with TransDecoder (http://transdecoder.github.io/). For each assembled component, only the longest ORF was retained, and gene families were delimited using all against all BLASTP (Altschul et al. 1990) and SiLiX (Miele et al. 2011). SiLiX parameters were set to $i = 0.6$ and $r = 0.6$ as they were maximizing the number of 1-to-1 orthologous gene families.

d_N / d_S calculation

Single-copy orthologs were extracted for three different sets of taxa: the 11 asellid species pairs (320 genes), the ibero-aquitanian clade (863 genes), and the alpine-coxalis clade (2257 genes) (Fig. 1). Each gene family was then aligned with the following procedure: (1) search and masking of frameshift using MACSE with frameshift cost set to −10 (Ranwez et al. 2011); (2) multiple alignments of the translated sequences with PRANK (Löytynoja and Goldman 2008) using the empirical codon model and F option; (3) site masking with Gblocks (Castresana 2000) using -t = c, -b5 = h and -b2 set as -b1. After gene concatenation, a transcriptome-wide d_N/d_S ratio was estimated with the free ratio model of CODEML from the package PAML 4.7a (Yang 2007). Confidence intervals were obtained using 100 nonparametric bootstrap samples (random sampling of the codon sites with replacement).

To test whether the observed d_N/d_S increase could be attributed to a reduction in the efficacy of purifying selection or to an increase in positive selection (a higher number of sites under positive selection and/or an elevated positive selection intensity), we used the M10 branch-site model (Yang et al. 2000) as implemented in BppML (Dutheil and Boussau 2008). To reduce computation times, the analysis was performed using quartets of taxa: the two species of a pair and two additional surface species used to root the tree. Large alignments (>300,000 codons) were reduced by randomly sampling 280,000 codons, and only sites that were complete were retained. Purifying and positive selection intensity (w^- and w^+) and frequency [$fq(w^-)$ and $1 - fq(w^-)$] were estimated using the a posteriori mean site d_N/d_S using BppMixedLikelihoods (Dutheil and Boussau 2008).

Single-nucleotide polymorphism

Estimating population polymorphism from pooled RNA-seq samples is complicated by the fact that (1) RNA-seq is prone to both RT-PCR and sequencing errors (Gout et al. 2013); (2) polymorphism can be overestimated by hidden paralogs (Gayral et al. 2013); and (3) it is difficult to differentiate low frequency alleles from sequencing errors in pooled data sets (Futschik and Schlötterer 2010). Although an accurate estimate is currently out of reach, it is possible to obtain polymorphism estimates that are comparable across taxa. We developed a statistical design that (1) is conservative in defining polymorphism; (2) balances the sampling effort so that estimates obtained within a species pair are comparable; and (3) maximizes the number of analyzed genes to gain statistical power. Single-nucleotide polymorphism (SNP) was searched on a set of 5027 gene families that were present as a single copy in at least six of the 11 species pairs. Gene famillies with hidden paralogs were filtered by using 10× coverage DNA-seq data available for four species (*P. karamani, P. hercegovinenis, P. ibericus*, and *P. arthrodilus*). Gene families that had a DNA-seq coverage higher than the 90th percentile in any of these four species were filtered for any further polymorphim analysis. RNA-seq reads were aligned on the assembled ORF using BWA (aln algorithm) (Li and Durbin 2009). SAMtools (Li et al. 2009) was then used to generate a BAM file, discard duplicated reads, and export a BCF file. SNPs were filtered and called with BCFtools and vcfutils.pl with the following conservative filtering parameters: minimum read depth of 10, minimum number of reads supporting an allele of 4, and minimum distance to a gap set to 15. Then, SNPs were classified as synonymous or nonsynonymous. Only the genes with high coverage in both species of a pair (average coverage >50×; the same results were obtained with a lower coverage cutoff) were further considered to compute transcriptome-wide summary statistics (Supplemental Table S3). This design ensured that synonymous and nonsynonymous polymorphism estimates could be compared across taxa, although each of these estimates might be over- or underestimated. We then calculated the population mutation rate (θ) using the Watterson estimator:

$$\hat{\theta}_w = \frac{p_S}{\sum_{i=1}^{2n-1} \frac{1}{i}}$$

with p_S the frequency of synonymous segregating sites, and n the number of pooled individuals. Finally, we calculated the ratio of nonsynonymous over synonymous segregating sites p_N/p_S. Confidence intervals were obtained by bootstrapping the genes 10,000 times.

This polymorphism data set was also used to measure the direction of selection statistics (DoS) (Stoletzki and Eyre-Walker 2011) using:

$$DoS = \frac{D_n}{D_n + D_s} - \frac{P_n}{P_n + P_s}$$

with D_s and D_n the number of synonymous and nonsynonymous divergences, and P_s and P_n the number of synonymous and nonsynonymous polymorphisms. Divergences were measured with PAML 4.7a (Yang 2007), and polymorphisms were measured using the above described pipeline. *DoS* were measured gene by gene for every species pair and compared for every pair using a Wilcoxon signed rank test or globally using the median *DoS* per species.

Rate of adaptation and Tajima's *D*

For two species pairs (*P. beticus*, *P. jalionacus*, *P. coiffaiti*, and *P. cavaticus*), we performed additional 50-base single-end Illumina RNA-seq, but this time independently sequencing 4–5 individuals per species, allowing the estimation of allele frequencies. Reads were mapped on the assembled transcriptomes using BWA (mem algorithm) (Li and Durbin 2009), and SNPs were called using Reads2snp (Gayral et al. 2013). The site frequency spectra (SFS) were then unfolded using the alignment with the respective sister species orthologs. Only sites with nonambiguously reconstructed ancestral and derived allele were kept. We then used the Messer and Petrov approach (Messer and Petrov 2013) to directly estimate the proportion of adaptive substitutions (α) from the unfolded SFS. Confidence intervals for α were obtained by bootstrapping the SNP 1000 times. The same data set was also used to test if populations were at the mutation-drift equilibrium using Tajima's *D* test (Tajima 1989).

Colonization time

For 19 species, we were able to determine the sequence of one opsin gene. Sequences were determined using (1) transcriptome sequences, (2) Sanger sequencing using PCR primers (LWF1a and Scylla) and PCR conditions from Taylor et al. (2005), and (3) genomic Illumina reads as detailed below. For the latter, reads were mapped on the closest available opsin sequence following the same approach as for the SNP search, and a consensus was called with the SAMtools program suite.

To estimate colonization time, we used the loss of function observed in several subterranean species and postulated that the opsin gene loss of function took place at the time of groundwater colonization. We used a model with two d_N/d_S ratios, one for the functional opsins (ω_{surf}) and one for the nonfunctional opsin (ω_1), which was set to 1. We then defined the d_N/d_S of a branch leading to a subterranean taxa (ω_{subt}) as the weighted mean between these two defined ratios, such as:

$$\omega_{subt} = \omega_{surf}\frac{T-t}{T} + \omega_1\frac{t}{T}$$

with T the speciation time, and t the time of colonization. From this, we estimated the time of colonization as follows:

$$t = T \times \frac{\omega_{subt} - \omega_{surf}}{1 - \omega_{surf}}.$$

Another relevant parameter is the proportion of time a species has been subterranean since the divergence with its sister species, named the relative colonization time (*RCT*), which we estimated using:

$$RCT = \frac{t}{T} = \frac{\omega_{subt} - \omega_{surf}}{1 - \omega_{surf}}.$$

Opsin d_N/d_S was estimated using PAML free-ratio branch model, and the ω_{surf} was estimated as the average d_N/d_S of the surface species showing the most obvious surface phenotypes (*P. coiffaiti, P. coxalis, P. karamani, P. ibericus, P. meridianus*, and *P. beticus*).

Measurement of genome size

We measured GS for 41 species of asellid (including the 11 species pairs), four Atyidae (Pancrustacea, Decapoda), and two Rissoidae (Mollusca, Gastropoda) (Supplemental Table S6). After sampling, individuals were preserved at ambient temperature in silica gel. Measurements were conducted according to Vieira et al. (2002). Nuclei were extracted from the head of organisms (or from entire individuals when body size was less than 3 mm in length). Heads were crushed in 200 µL of cold modified Galbraith's nuclei isolation buffer (20 mM MOPS, 20.5 mM MgCl2, 35.5 mM trisodium citrate, 0.1% Triton X-100, 20 µg mL^{-1} boiled RNase A, pH 7.2 adjusted with NaOH) (Galbraith et al. 1983). The mixture was filtered through 100 µm and then 30 µm mesh-size nets. The filtrate was centrifuged for 10 sec at 2600*g*, and the supernatant was carefully removed. Pellets were resuspended in 200 µL of nuclei isolation buffer. The resuspension was again centrifuged for 10 sec at 2600*g*, and the supernatant was carefully removed. Pellets were resuspended in 250 µL of buffer and transferred to 5 mL polystyrene round-bottom tubes. An amount of 50 µL of propidium iodide was added to each tube. Tubes were kept in ice and darkness until GS measurements.

Genome sizes were measured using FACSCanto II flow cytometer (Becton Dickinson Instruments) fitted with an argon laser at 488-nm wavelength. We analyzed five individuals per species. Individuals were measured in a random order, and two individuals of the same species were never analyzed in the same run. Samples were calibrated to two external standards: *Drosophila virilis* females (GS of 0.41 pg) (Bosco et al. 2007) and *Asellus aquaticus* (GS of 2.49 pg) (Rocchi et al. 1988, and authors' cross validation). The *Drosophila* were maintained under laboratory conditions at 25°C for two to three generations before GS measurements. Standards were prepared using the protocol described above from five organism heads and were measured in each run (two measurements of *D. virilis* at the beginning and end of runs and five measurements of *A. aquaticus* evenly distributed during the runs).

The FlowQ bioconductor package (Gentleman et al. 2004) in R software (R Core Team 2013) was used for quality assessment of flow cytometry data. All the flow cytometer analyses were checked for cell number, boundary events, and time anomalies. Cell subsetting known as gating, was first performed manually using the BD FACSDiva software (BD Biosciences). Second, the automatic curvHDR filtering method (Naumann et al. 2010) was used to select the cells located in the highest density region (HDR level = 0.8). Then, when drift over time was significant, gated values were corrected using a linear regression on *A. aquaticus* reference using the following equation:

$$IP_{coor} = IP_x - \frac{IP_x}{IP_{st}} \times (x - x_0) \times \lambda,$$

where IP_{corr} is the corrected gated value, IP_x is the gated value to correct, IP_{st} is the *A. aquaticus* reference gated value at the beginning of the run (time $t = x_0$) estimated by the linear regression, x is the measurement time for the gated value to correct, x_0 is the time at the beginning of the run, and λ is the slope of the linear regression on *A. aquaticus* reference. Drift was considered significant when the regression on *A. aquaticus* reference had adjusted R^2 values ≥0.1. Finally, GS was derived from fluorescence data using *D. virilis* as a standard for Asellidae and Rissoidae and using *A. aquaticus* as a standard for Atyidae. Indeed, large GS in Atyidae (previously known GS range from 3.30 to 7.20 pg) (http://www.genomesize.com/) prevented the use of *D. virilis* as a standard.

Growth rate and body size

Growth rates were estimated using the total RNA normalized by the total protein biomass of an organism (RNA/protein ratio, Wojewodzic et al. 2011) for at least seven individuals per species. Total RNA and proteins were isolated using TRI Reagent (Molecular Research Center). RNA concentrations were calculated by fluorometry with a Qubit (Life Technologies). Total proteins were obtained using the Bicinchoninic acid assay (Smith et al. 1985). Body size was estimated using maximum body size as reported in each species description.

Genome sequencing

To compare the size of the repeatome between surface and subterranean species, we sequenced the genome of the 11 pairs of Asellidae species. For four species (*P. ibericus, P. arthrodilus, P. karamani, P. hercegovinensis*), we built blunt-ended libraries for shotgun sequencing on Illumina platforms, as described (Orlando et al. 2013; Seguin-Orlando et al. 2013) with few modifications. One microgram DNA in 100 µL TE buffer was sheared using a Bioruptor NGS device (Diagenode) with four cycles of 15 sec ON/90 sec OFF. The obtained size distributions of sheared DNA fragments were centered at around 500 bp. After concentration in 22 µL EB buffer (Qiagen) with the MinElute PCR Purification kit (Qiagen), the sheared DNA fragments were built into blunt-ended DNA libraries using the NEBNext Quick DNA Library Prep Master Mix Set for 454 (New England BioLabs, reference E6070L), following the protocol described previously (Meyer and Kircher 2010), but with 0.5 µM Illumina adapters (final concentration). All reactions were carried out in 25 µL volumes; incubation times and temperatures were as follows: 20 min at 12°C, 15 min at 37°C for end-repair; 20 min at 20°C for ligation; 20 min at 37°C, 20 min at 80°C for fill-in. After the end-repair and ligation steps, reaction mixes were purified using the MinElute PCR Purification Kit (Qiagen) using elution volumes of 16 µL and 22 µL of EB buffer, respectively. The final 25 µL volume of blunt-end libraries was split in two parts and PCR amplified independently in 50 µL reaction mixes containing: 5 units Taq Gold (Life Technologies), 1× Gold Buffer, 4 mM MgCl2, 1 mg/mL BSA, 0.25 mM of each dNTP, 0.5 µM of primer PE1.0 (5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′), and 0.5 µM of an Illumina 6 bp-indexed ("I") primer (5′-CAAGCAGAAGACGGCATACGAGATIIIIIIGTGACTGGAGTTCAGACGTGTGCTCTTCCG-3′). Thermocycling conditions for the amplifications were: activation for 10 min at 92°C; followed by nine cycles of denaturation for 30 sec at 92°C, annealing for 30 sec at 60°C, elongation for 30 sec at 72°C; and final elongation for 7 min at 72°C. PCR products were purified using the MinElute PCR Purification kit, with a final elution volume of 25 µL EB buffer.

For the remaining 18 species, we built Illumina TruSeq DNA PCR-free LT libraries (Illumina, catalog FC-121-3001), following manufacturer's recommendations. Briefly, 1 µg of DNA extract was sheared in a total volume of 50 µL TE buffer, using a Bioruptor NGS device (Diagenode) with three cycles of 25 sec ON/90 sec OFF. The fragmented DNA was cleaned up using Illumina Sample Purification Beads. After end repair, the DNA fragments were size-selected around 350 bp using two consecutive bead purification steps, A-tailed, and ligated to 6-bp indexed Illumina TruSeq adapters (Set A). Two last bead purifications were performed to remove any adapter dimer, and the final libraries were resuspended in a volume of 20 µL Resuspension Buffer. In order to control for contamination, library and PCR blanks were carried out at the same time as the samples. Amplified libraries and blanks were quantified using the 2100 Bioanalyzer (Agilent) High-Sensitivity DNA Assay. No detectable amount of DNA could be recovered from the blanks. Blunt-End indexed DNA libraries were pooled and sequenced on two lanes of a HiSeq 2000 Illumina platform (100 cycles paired-end mode run), and the two PCR-free library pools were each sequenced on one flow cell of the HiSeq 2500 Illumina platform (150 paired-end run, Rapid Mode, 6-bp index read), at the Danish National High-Throughput DNA Sequencing Centre.

Repeatome size estimates

We used low coverage read sequencing to characterize repetitive genome sequences (Novák et al. 2010) using RepeatExplorer (Novák et al. 2013). Prior to analysis, DNA-seq reads were randomly sampled to achieve 0.05× coverage following the GS estimated by flow cytometry, so that estimates are comparable across taxa. After clustering of the reads into highly repetitive elements by RepeatExplorer, the number of reads representing each repeat element is a direct function of the repeat frequency, the GS, and the sequencing effort. The proportion of the genome (GP) composed of this repeat is $GP_i = n_{reads_i}/n_{reads}$, with n_{reads_i} the number of reads mapping to the repeat i, and n_{reads} the total number of reads. The proportion of genome composed of repeats (the repeatome size) is then $GP = \sum_i GP_i$. By default, RepeatExplorer filters repeat elements that have genome proportion <0.01%. To achieve comparable estimates independent of genome sizes, the number of repeats or the repeatome size of a given genome was recalculated by filtering repeats that occupied <0.5 Mb. A repeat element total genomic size (TGS) was then calculated as $TGS_i = GP_i \times GS_j$ with GS_j the genome size of species j. Repeat families were delimited using blastn (E-value = 0.1) and SiLiX ($i = 50$ and $r = 70$).

Comparative analyses

Phylogenetic generalized least-squares (PGLS) regression models (Martins and Hansen 1997) were used to test for the correlation between two variables. We first tested the association between the ecological transition and population size, biological traits, or GS (Table 1, top). PGLS were also used to test for the association between GS and population size, biological traits (Table 1, bottom) or genomic features (Table 2). The correlation between two variables was assessed by comparing a model without the predictor variable (intercept model) to a model including the predictor variable using a likelihood ratio test (LRT). Analyses were performed in R using APE (Paradis et al. 2004) and nlme (https://cran.r-project.org/package=nlme) packages. The best model of trait evolution and its associated covariance structure—in our case, the Brownian motion model—was selected according to minimum Akaike information criterion (AIC). The difference in $\hat{\theta}_w$, p_N/p_S, and d_N/d_S between the two species of a surface–subterranean pair was tested using the proportion of bootstrap replicates supporting a difference (critical level = 5%), and the difference in GS was tested using a Wilcoxon rank-sum test with five measurements of GS (i.e., five individuals) per species. We also performed ordinary least-squares models to test for the effect of the ecological status on GS while ignoring phylogenetic relationships among species. To test for the effect of the ecological transition on GS using a wider range of taxa (i.e., 18 species pairs including Decapoda and Mollusca with five measurements of GS per species), we performed a linear mixed model in R using the nlme package because a chronogram with accurate branch length could not be obtained given the available molecular markers and calibration points. The ecological status (i.e., surface versus subterranean) was a fixed effect, and we specified the random error structure as ecological status nested into species pairs to account for phylogenetic relationships among species. Then, we performed the model with no hierarchy in the random error structure, which is equivalent to an ordinary least-squares model, to test for the effect of the ecological status on GS while ignoring phylogenetic relationships among species. Differences in the coefficients of variation of different variables were tested using modified signed-likelihood ratio test (Krishnamoorthy and Lee 2014) using the R package cvequality.

Data access

Sequence reads and assemblies from this study have been submitted to the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena) under study accession number PRJEB14193. Sanger sequences from this study have been submitted to NCBI GenBank (https://www.ncbi.nlm.nih.gov/nucleotide/) under accession numbers KC610091–KC610505 (Supplemental Table S9).

Acknowledgments

This work was supported by the Agence Nationale de la Recherche (ANR-08-JCJC-0120-01 DEEP, ANR-15-CE32-0005 Convergenomix); the Institut Universitaire de France; the European Commission (7th EU Framework Programme, Contract No. 226874, BioFresh), the CNRS (APEGE No. 70632, PEPS ExoMod 2014 and Enviromics 2014), the French Embassy in Denmark, the Danish National Research Foundation (DNRF94), and the Marie-Curie Intra-European Fellowship (IEF 302617). We thank the following collaborators and organizations for their help in collecting biological material: Alhama de Granada municipality, J.P. Besson, M. Bodon, M. Bouillon, C. Capderrey, I. Carreira, B. Chatelier, C. Colson-Proch, M. Creuzé des Chatelliers, Cueva de Valporquero, D. Culver, T. Datry, Delegacion provincial de medio ambiente Ronda, J. Ferrandini, M. Ferrandini, C. Fiser, D. Fong, D. Gina, S. Gottstein, Gouffre de Padirac, B. Kaufmann, L. Knight, E. Lana, R. Le Pennec, F. Lescher-Moutoue, G. Magniez, G. Messana, G. Michel, E. Nassar, N. Nassar-Simon, J. Notenboom, R. Orozco, Parc National du Mercantour, Parco Naturale Alta Valle Pesio e Tanaro, S. Planes, V. Prié, A.S. Reboleira, Sauve municipality, A. Sendra, L. Simon, B. Sket, F. Stoch, M. Turjak, and M. Zagmajster. We thank the laboratory technicians at the Danish High-Throughput DNA Sequencing Centre for technical assistance. We thank M. Boulesteix, N. Burley, C. Goubert, E. Paradis, and L. Simon for providing advice on the analyses; C. Biémont and S. Schaack for their advice or comments on earlier drafts of this manuscript. We gratefully acknowledge support from the CNRS/IN2P3 Computing Center (Lyon/Villeurbanne, France), for providing a significant amount of the computing resources needed for this work. This work was performed using the computing facilities of the CC LBBE/PRABI. We thank two anonymous reviewers and Adam Eyre-Walker for their comments.

References

Ai B, Wang ZS, Ge S. 2012. Genome size is not correlated with effective population size in the *oryza* species. *Evolution* **66:** 3302–3310.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403–410.

Arnqvist G, Sayadi A, Immonen E, Hotzy C, Rankin D, Tuda M, Hjelmen CE, Johnston JS. 2015. Genome size correlates with reproductive fitness in seed beetles. *Proc R Soc Lond B Biol Sci* **282:** 20151421.

Aronesty E. 2013. Comparison of sequencing utility programs. *Open Bioinforma J* **7:** 1–8.

Bosco G, Campbell P, Leiva-Neto JT, Markow TA. 2007. Analysis of drosophila species genome size and satellite DNA content reveals significant differences among strains as well as between species. *Genetics* **177:** 1277–1290.

Calvignac S, Konecny L, Malard F, Douady CJ. 2011. Preventing the pollution of mitochondrial datasets with nuclear mitochondrial paralogs (*numts*). *Mitochondrion* **11:** 246–254.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17:** 540–552.

Cavalier-Smith T. 1982. Skeletal DNA and the evolution of genome size. *Annu Rev Biophys Bioeng* **11:** 273–302.

Charlesworth B, Barton N. 2004. Genome size: does bigger mean worse? *Curr Biol* **14:** R233–R235.

Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284:** 601–603.

Dutheil J, Boussau B. 2008. Non-homogeneous models of sequence evolution in the bio++ suite of libraries and programs. *BMC Evol Biol* **8:** 255.

Emerling CA, Springer MS. 2014. Eyes underground: regression of visual protein networks in subterranean mammals. *Mol Phylogenet Evol* **78:** 260–270.

Fang X, Nevo E, Han L, Levanon EY, Zhao J, Avivi A, Larkin D, Jiang X, Feranchuk S, Zhu Y, et al. 2014. Genome-wide adaptive complexes to underground stresses in blind mole rats *Spalax*. *Nat Commun* **5:** 3966.

Fierst JL, Willis JH, Thomas CG, Wang W, Reynolds RM, Ahearne TE, Cutter AD, Phillips PC. 2015. Reproductive mode and the evolution of genome size and structure in *Caenorhabditis* nematodes. *PLoS Genet* **11:** e1005323.

Futschik A, Schlötterer C. 2010. The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics* **186:** 207–218.

Galbraith DW, Harkins KR, Maddox JM, Ayres NM, Sharma DP, Firoozabady E. 1983. Rapid flow cytometric analysis of the cell cycle in intact plant tissues. *Science* **220:** 1049–1051.

Galtier N. 2016. Adaptive protein evolution in animals and the effective population size hypothesis. *PLoS Genet* **12:** 1–23.

Gayral P, Melo-Ferreira J, Glémin S, Bierne N, Carneiro M, Nabholz B, Lourenco JM, Alves PC, Ballenghien M, Faivre N, et al. 2013. Reference-free population genomics from next-generation transcriptome data and the vertebrate–invertebrate gap. *PLoS Genet* **9:** e1003457.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5:** R80.

Gout JF, Thomas WK, Smith Z, Okamoto K, Lynch M. 2013. Large-scale detection of in vivo transcription errors. *Proc Natl Acad Sci* **110:** 18584–18589.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29:** 644–652.

Gregory TR. 2001. Coincidence, coevolution, or causation? DNA content, cellsize, and the C-value enigma. *Biol Rev* **76:** 65–101.

Gregory TR. 2005. *The evolution of the genome*. Elsevier, San Diego, CA.

Gregory TR, Witt JD. 2008. Population size and genome size in fishes: a closer look. *Genome* **51:** 309–313.

Grime J, Mowforth M. 1982. Variation in genome size—an ecological interpretation. *Nature* **299:** 151–153.

Huntsman BM, Venarsky MP, Benstead JP, Huryn AD. 2011. Effects of organic matter availability on the life history and production of a top vertebrate predator (plethodontidae: *Gyrinophilus palleucus*) in two cave streams. *Freshwater Biol* **56:** 1746–1760.

James JE, Piganeau G, Eyre-Walker A. 2016. The rate of adaptive evolution in animal mitochondria. *Mol Ecol* **25:** 67–78.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484:** 55–61.

Krishnamoorthy K, Lee M. 2014. Improved tests for the equality of normal coefficients of variation. *Comput Stat* **29:** 215–232.

Kuo CH, Moran NA, Ochman H. 2009. The consequences of genetic drift for bacterial genome complexity. *Genome Res* **19:** 1450–1454.

Leys R, Cooper SJ, Strecker U, Wilkens H. 2005. Regressive evolution of an eye pigment gene in independently evolved eyeless subterranean diving beetles. *Biol Lett* **1:** 496–499.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25:** 1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079.

Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320:** 1632–1635.

Lynch M. 2007. *The origins of genome architecture*, Vol. 98. Sinauer, Sunderland, MA.

Lynch M. 2011. Statistical inference on the mechanisms of genome evolution. *PLoS Genet* **7:** e1001389.

Lynch M, Conery JS. 2003. The origins of genome complexity. *Science* **302:** 1401–1404.

Lynch M, Bobay LM, Catania F, Gout JF, Rho M. 2011. The repatterning of eukaryotic genomes by random genetic drift. *Annu Rev Genomics Hum Genet* **12:** 347–366.

Martins EP, Hansen TF. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am Nat* **149:** 646–667.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351:** 652–654.

Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald-Kreitman test. *Proc Natl Acad Sci* **110:** 8615–8620.

Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc* **2010:** pdb.prot5448.

Miele V, Penel S, Duret L. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* **12:** 116.

Mohlhenrich ER, Mueller RL. 2016. Genetic drift and mutational hazard in the evolution of salamander genomic gigantism. *Evolution* **70:** 2865–2878.

Morvan C, Malard F, Paradis E, Lefébure T, Konecny-Dupré L, Douady C. 2013. Timetree of aselloidea reveals species diversification dynamics in groundwater. *Syst Biol* **62:** 512–522.

Naumann U, Luta G, Wand MP. 2010. The curvHDR method for gating flow cytometry samples. *BMC Bioinformatics* **11:** 44.

Nielsen R, Yang Z. 2003. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol Biol Evol* **20:** 1231–1239.

Niemiller ML, Fitzpatrick BM, Shah P, Schmitz L, Near TJ. 2013. Evidence for repeated loss of selective constraint in rhodopsin of amblyopsid cavefishes (teleostei: *Amblyopsidae*). *Evolution* **67:** 732–748.

Novák P, Neumann P, Macas J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* **11:** 378.

Novák P, Neumann P, Pech J, Steinhaisl J, Macas J. 2013. RepeatExplorer: a galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29:** 792–793.

Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annu Rev Ecol Syst* **23:** 263–286.

Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E, Petersen B, Moltke I, et al. 2013. Recalibrating *Equus* evolution using the genome sequence of an early middle pleistocene horse. *Nature* **499:** 74–78.

Otto SP. 2007. The evolutionary consequences of polyploidy. *Cell* **131:** 452–462.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20:** 289–290.

Petrov DA. 2001. Evolution of genome size: new approaches to an old problem. *Trends Genet* **17:** 23–28.

Protas ME, Hersey C, Kochanek D, Zhou Y, Wilkens H, Jeffery WR, Zon LI, Borowsky R, Tabin CJ. 2005. Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. *Nat Genet* **38:** 107–111.

Purvis A, Gittleman JL, Luh HK. 1994. Truth or consequences: effects of phylogenetic accuracy on two comparative methods. *J Theor Biol* **167:** 293–300.

R Core Team. 2013. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Ranwez V, Harispe S, Delsuc F, Douzery EJ. 2011. MACSE: multiple alignment of coding sequences accounting for frameshifts and stop codons. *PLoS One* **6:** e22594.

Rocchi A, Lanza V, Di Castro M. 1988. Surface spreading of synaptonemal complexes in three isopod crustacean species. *Genetica* **78:** 125–132.

Seguin-Orlando A, Schubert M, Clary J, Stagegaard J, Alberdi MT, Prado JL, Prieto A, Willerslev E, Orlando L. 2013. Ligation bias in Illumina next-generation DNA libraries: implications for sequencing ancient genomes. *PLoS One* **8:** e78575.

Smith P, Krohn RI, Hermanson G, Mallia A, Gartner F, Provenzano M, Fujimoto E, Goeke N, Olson B, Klenk D. 1985. Measurement of protein using bicinchoninic acid. *Anal Biochem* **150:** 76–85.

Soria-Carrasco V, Gompert Z, Comeault AA, Farkas TE, Parchman TL, Johnston JS, Buerkle CA, Feder JL, Bast J, Schwander T, et al. 2014. Stick insect genomes reveal natural selection's role in parallel speciation. *Science* **344:** 738–742.

Stoletzki N, Eyre-Walker A. 2011. Estimation of the neutrality index. *Mol Biol Evol* **28:** 63–70.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585–595.

Taylor SD, de la Cruz KD, Porter ML, Whiting MF. 2005. Characterization of the long-wavelength opsin from mecoptera and siphonaptera: Does a flea see? *Mol Biol Evol* **22:** 1165–1174.

Thomas C. 1971. The genetic organization of chromosomes. *Annu Rev Genet* **5:** 237–256.

Venarsky MP, Huntsman BM, Huryn AD, Benstead JP, Kuhajda BR. 2014. Quantitative food web analysis supports the energy-limitation hypothesis in cave stream ecosystems. *Oecologia* **176:** 859–869.

Vieira C, Nardon C, Arpin C, Lepetit D, Biémont C. 2002. Evolution of genome size in *Drosophila* is the invader's genome being invaded by transposable elements? *Mol Biol Evol* **19:** 1154–1161.

Vinogradov AE. 1995. Nucleotypic effect in homeotherms: body-mass-corrected basal metabolic rate of mammals is related to genome size. *Evolution* **49:** 1249–1259.

Whitney KD, Garland T. 2010. Did genetic drift drive increases in genome complexity? *PLoS Genet* **6:** e1001080.

Whitney KD, Baack EJ, Hamrick JL, Godt MJW, Barringer BC, Bennett MD, Eckert CG, Goodwillie C, Kalisz S, Leitch IJ, et al. 2010. A role for non-adaptive processes in plant genome size evolution? *Evolution* **64:** 2097–2109.

Whitney KD, Boussau B, Baack EJ, Garland T. 2011. Drift and genome complexity revisited. *PLoS Genet* **7:** e1002092.

Wojewodzic MW, Rachamim T, Andersen T, Leinaas HP, Hessen DO. 2011. Effect of temperature and dietary elemental composition on RNA/protein ratio in a rotifer. *Funct Ecol* **25:** 1154–1160.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24:** 1586–1591.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155:** 431–449.

Yang L, Takuno S, Waters ER, Gaut BS. 2011. Lowly expressed genes in *Arabidopsis thaliana* bear the signature of possible pseudogenization by promoter degradation. *Mol Biol Evol* **28:** 1193–1203.

Yi S, Streelman JT. 2005. Genome size is negatively correlated with effective population size in ray-finned fish. *Trends Genet* **21:** 643–646.

Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu SH. 2009. Evolutionary and expression signatures of pseudogenes in arabidopsis and rice. *Plant Physiol* **151:** 3–15.

Received July 8, 2016; accepted in revised form March 30, 2017.

Less effective selection leads to larger genomes

Tristan Lefébure, Claire Morvan, Florian Malard, et al.

Genome Res. 2017 27: 1016-1028 originally published online April 19, 2017
Access the most recent version at doi:10.1101/gr.212589.116